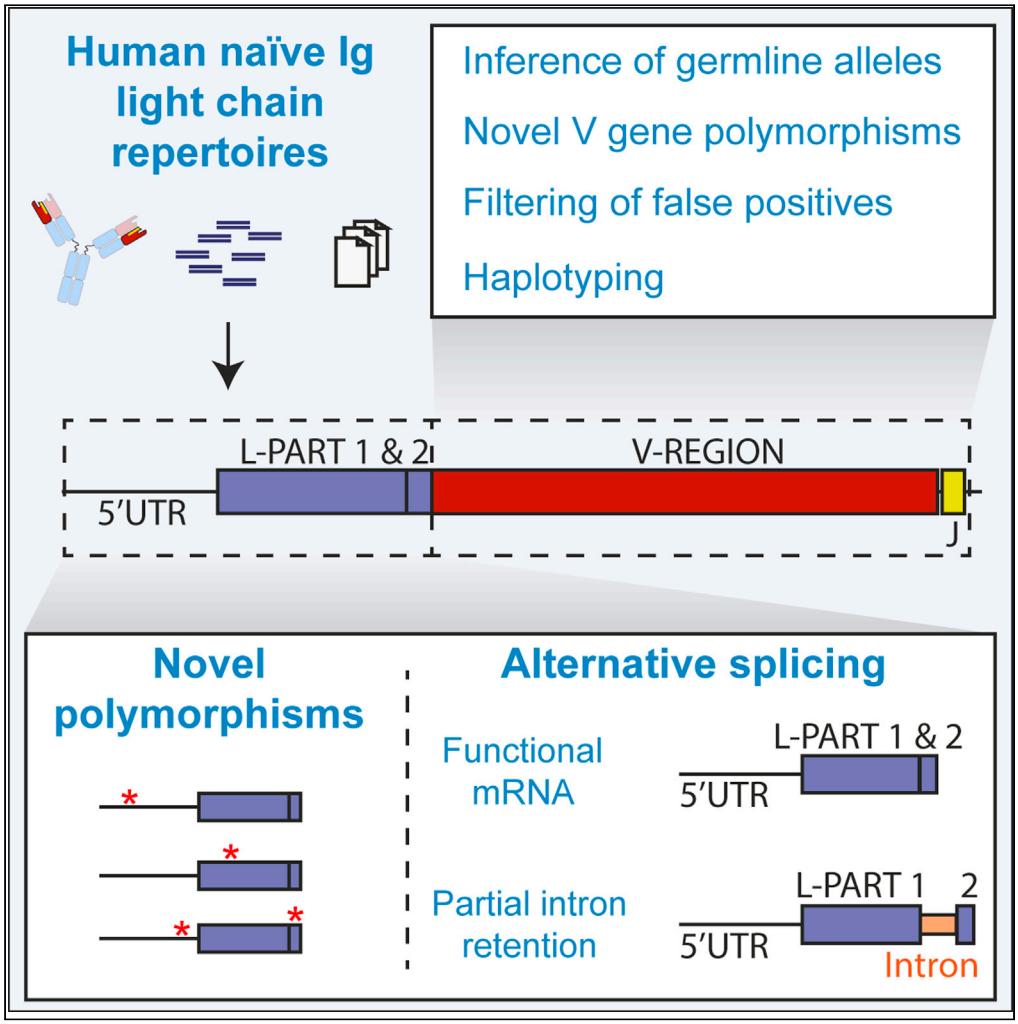
iScience

CellPress
OPEN ACCESS

Article

Germline polymorphisms and alternative splicing of human immunoglobulin light chain genes



Ivana Mikocziova, Ayelet Peres, Moriah Gidoni, Victor Greiff, Gur Yaari, Ludvig M. Sollid

l.m.sollid@medisin.uio.no

Highlights

Inference of germline light chain Ig V alleles and filtering of sequencing artefacts

48 previously unreported V gene alleles, 25 in kappa and 23 in lambda genes

Partial intron detected in some upstream sequences suggesting alternative splicing

Novel polymorphisms in the upstream regions: 5'UTR, L-PART1, L-PART2

Mikocziova et al., iScience *24*, 103192
October 22, 2021
https://doi.org/10.1016/j.isci.2021.103192

iScience

CellPress
OPEN ACCESS

Article

Germline polymorphisms and alternative splicing of human immunoglobulin light chain genes

Ivana Mikocziova,[1,2,5] Ayelet Peres,[3,4,5] Moriah Gidoni,[3] Victor Greiff,[2] Gur Yaari,[3,4,6] and Ludvig M. Sollid[1,2,6,7,*]

SUMMARY

Inference of germline polymorphisms in immunoglobulin genes from B cell receptor repertoires is complicated by somatic hypermutations, sequencing/PCR errors, and by varying length of reference alleles. The light chain inference is particularly challenging owing to large gene duplications and absence of D genes. We analyzed the light chain cDNA sequences from naïve B cell receptor repertoires from 100 individuals. We optimized light chain allele inference by tweaking parameters of the TIgGER functions, extending the germline reference sequences, and establishing mismatch frequency patterns at polymorphic positions to filter out false-positive candidates. We identified 48 previously unreported variants of light chain variable genes. We selected 14 variants for validation and successfully validated 11 by Sanger sequencing. Clustering of light chain 5′UTR, L-PART1, and L-PART2 revealed partial intron retention in 11 kappa and 9 lambda V alleles. Our results provide insight into germline variation in human light chain immunoglobulin loci.

INTRODUCTION

Immunoglobulins (Igs) are essential molecules of the immune system that can recognize and bind a variety of antigens. They are produced by B cells and can be secreted as antibodies or can be immobilized on the B cell surface in the form of a B cell receptor (BCR). Immunoglobulins are formed by two identical dimers, and each dimer contains one heavy chain paired with one light chain. These dimers are assembled in a structure that resembles the letter Y. The two "arms" of the Y letter-shaped antibody contain a paratope that interacts with an antigen (Akbar et al., 2021; Sela-Culang et al., 2013). This paratope is formed by the variable domains of the heavy and light chains. The variable domains are coded by a large number of variable (V) genes that can recombine with a number of diversity (D) and joining (J) genes within the same locus (Schatz, 2004). In the light chain, the D region is absent and therefore the V region recombines directly with the J region (Collins and Watson, 2018). In humans, kappa chain genes (IGK) are located on chromosome 2 (2p11.2) whereas lambda chain genes (IGL) are located on chromosome 22 (22q11.2) (McBride et al., 1982).

V(D)J recombination together with different heavy and light chain pairing options contributes to the large diversity of antibody paratopes, which enables the recognition of many different antigens. An additional level of diversity can be introduced during B cell maturation via a process called somatic hypermutation, which introduces mutations in the V genes to increase their binding affinity for an antigen (Chi et al., 2020). However, it is important to remember that germline variation also plays a great role in shaping an individual's antibody/BCR repertoire (Glanville et al., 2011; Rubelt et al., 2016; Slabodkin et al., 2021; Watson et al., 2017). Different germline variants of the same V gene can give rise to antibodies with slightly different amino acid sequence and different affinity to the same antigen (Avnir et al., 2014). Of importance, Ig germline variants have been demonstrated to affect susceptibility to infection (Avnir et al., 2016; Tan et al., 2018) and autoimmune diseases (Johnson et al., 2020; Vencovský et al., 2002) thus underscoring the need to comprehensively map Ig germline variation (Collins et al., 2020; Mikocziova et al., 2021).

Germline V gene variants can be inferred from BCR repertoire sequencing data by using specialized software (Corcoran et al., 2016; Gadala-Maria et al., 2019, 2015; Ralph and Matsen (IV), 2019). Despite the availability of different germline inference tools, the inference of Ig light chain genes from BCR repertoire data is not straightforward. Large gene duplications in the kappa locus make inference difficult. This is because most of the duplicated genes, despite lying on a different part of the locus, have identical nucleotide sequences (Watson et al., 2015). Software tools used for inference of immunoglobulin genes from repertoire

[1]K.G. Jebsen Centre for Coeliac Disease Research, Institute of Clinical Medicine, University of Oslo, 0372 Oslo, Norway

[2]Department of Immunology, Oslo University Hospital, 0372 Oslo, Norway

[3]Faculty of Engineering, Bar Ilan University, Ramat Gan 5290002, Israel

[4]Bar Ilan Institute of Nanotechnologies and Advanced Materials, Bar Ilan University, Ramat Gan 5290002, Israel

[5]These authors contributed equally

[6]These authors contributed equally

[7]Lead contact

*Correspondence: l.m.sollid@medisin.uio.no

https://doi.org/10.1016/j.isci.2021.103192

data (Corcoran et al., 2016; Gadala-Maria et al., 2019, 2015; Ralph and Matsen (IV), 2019) have been mostly used for the Ig heavy chain (Gadala-Maria et al., 2015; Luo et al., 2019; Scheepers et al., 2015; Thörnqvist and Ohlin, 2018), and studies looking into the germline variation in the human Ig light chain are currently lacking. The lack of attention given to the light chain genes, particularly in terms of germline variation, is concerning since mutational status of immunoglobulin genes is often used as a prognostic marker for different diseases such as chronic lymphocytic leukemia (Tobin, 2005). Biases in gene inference can lead to incorrect conclusions (Xochelli et al., 2015). Apart from that, the lack of knowledge about germline variation in immunoglobulin genes hinders progress and prevents us from exploiting this knowledge for the improvement of diagnostic and/or therapeutic methods.

In this study, we analyzed a dataset of naïve BCR repertoires from a Norwegian cohort of 100 individuals. Although the dataset was previously published, only the heavy chain has been analyzed so far (Gidoni et al., 2019; Mikocziova et al., 2020). We focused on the detection of germline variation in light chain genes, and our analysis revealed several previously unreported germline V gene polymorphisms. We provide an improved strategy for inferring light chain alleles from immunoglobulin repertoire data. In addition to adjusting TIgGER (Gadala-Maria et al, 2015, 2019) parameters, we have also exploited mismatch frequency of polymorphic position with a custom-set threshold derived from the population expression distribution, to help us identify false-positive candidates. Our approach was guided by targeted amplification and Sanger sequencing of genes with true positives as well as suspected false-positive candidates and borderline cases. Furthermore, we clustered the upstream sequences of the light chain transcripts, which revealed alternative splicing in certain kappa genes. Together, our data reveal substantial germline diversity in the kappa and lambda V genes, and we also show evidence that points to kappa gene expression being regulated via alternative splicing.

RESULTS

Optimization of allele inference for light chain immunoglobulin genes

One of the main challenges that complicate the annotation of kappa alleles is the large duplication event in the kappa locus. Since most of the duplicated genes share the same alleles and are virtually identical, it is impossible for the annotation software to decide which of the duplicated genes a sequence might come from. To prevent ambiguous allele annotation and for the purpose of this analysis, we treated each duplicated pair of genes that has at least one shared identical allele as one gene. As for the germline reference sequences, we collapsed identical sequences and added the letter "E" to their gene name in our customized reference database. For example, the germline reference sequences of IGKV1-12*01 and IGKV1D-12*01 are identical; therefore, we kept only one and named it IGKV1E-12*01 to indicate that it can originate from either of these genes.

The length of the germline reference alleles also affects the annotation and inference process. Several genes in the IMGT germline reference database are shorter in the 5′ and 3′ ends. This causes the annotation software to annotate our sequences based on their length rather than nucleotide sequence identity. To solve this problem, we artificially extended the reference with a consensus sequence from the other alleles of the gene.

Finally, the sequencing depth and the number of sequences for each gene also influence the process of inference. The number of Ig lambda (IGL) sequences is a third of the size of kappa (IGK). In addition, more mutations were present in the light chain V genes compared with the heavy chain, suggesting non-naive sequences. We tweaked the TIgGER function to infer for lower sequence depth. This resulted in many novel alleles inferred, which we filtered before genotyping. We only allowed up to two novel allele inferences for each gene that had the highest value of "novel_allele_count."

Identification of previously unreported polymorphisms in the light chain V genes

For the discovery of potential novel allele candidates in the light chain V genes, we used TIgGER (Gadala-Maria et al, 2015, 2019) and IgDiscover (Corcoran et al., 2016). The aim was not to compare these tools, but rather to add an additional level of confidence to our analysis. Both TIgGER and IgDiscover inferred several novel allele candidates in the light chain sequences of our data, and the overlap between these two software tools varied (Figure 1). Several candidates for novel polymorphisms were identified exclusively in four individuals that were sequenced in a separate batch as a pilot. These candidates were primarily found by

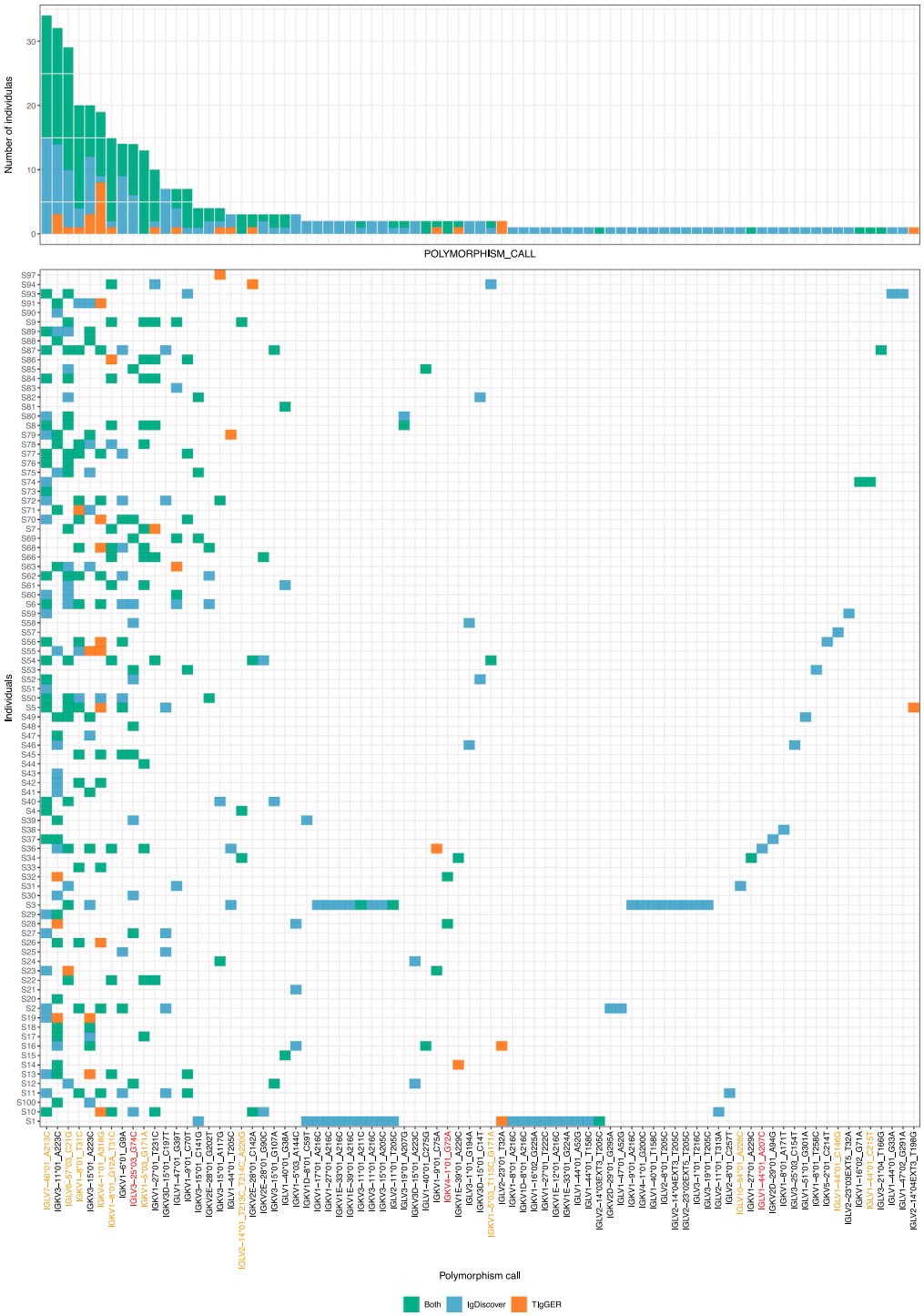


Figure 1. Novel light chain V gene polymorphism candidates

Novel alleles were inferred using IgDiscover and TIgGER software suites. For uniformity, personal genotypes of both Ig light chain loci were inferred using TIgGER. All novel alleles from both methods that were inferred at least once in the genotype with a confidence level higher than 10 (K) appear on the x axis. Alleles on the x axis that were validated by Sanger sequencing are marked in orange and those that were not validated despite attempts to do so are marked in red. For each allele, the color of a tile or bar represents the novel allele inference method that was used. The height of each bar on top represents the number of individuals for whom the allele appeared in the genotype.

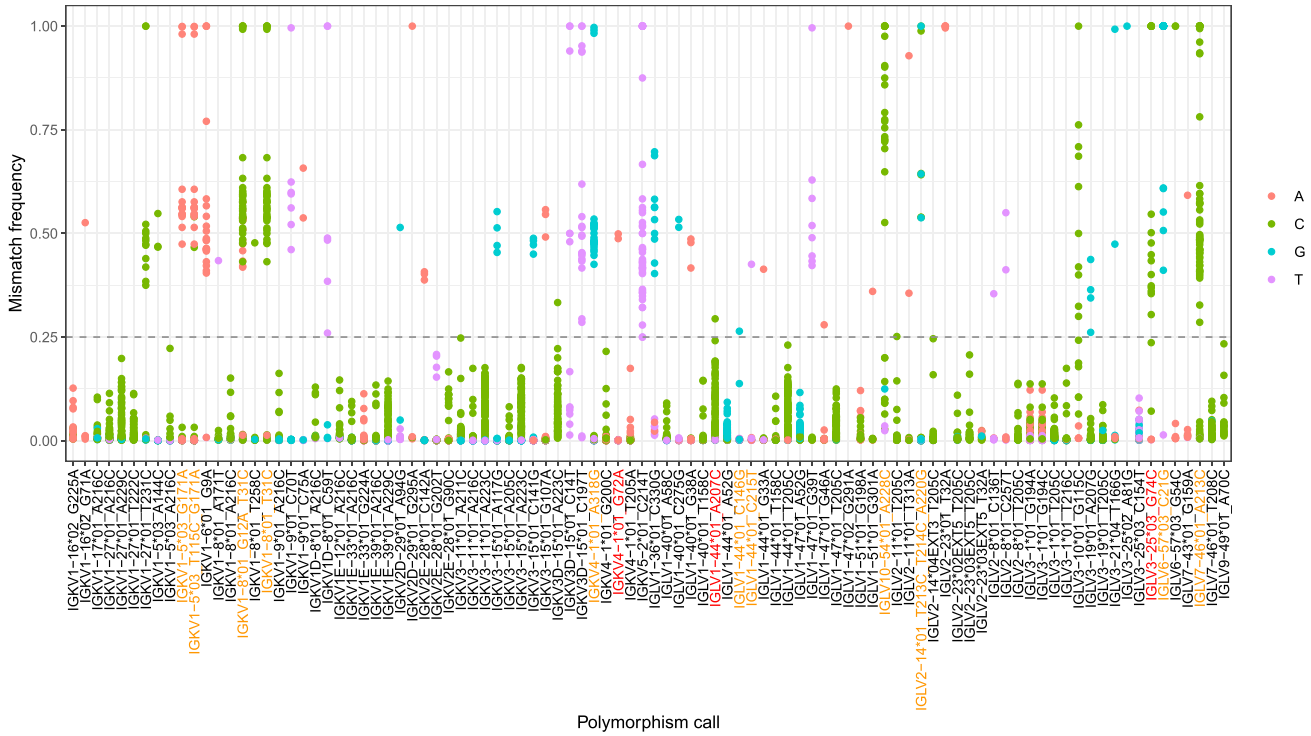


Figure 2. Mismatch frequency helps identify sequencing artefacts

The x axis shows novel allele candidates inferred by TIgGER. The relative mismatch frequency of each polymorphism is shown on the y axis. Each dot represents the mismatch frequency of a single individual for a certain allele. The colors of the dots represent the nucleotide that does not match the germline. Novel polymorphism candidates on the x axis that are colored in orange were validated by Sanger sequencing, and the red ones were attempted but not validated.

IgDiscover. This could be because we kept the IgDiscover parameters as default and only optimized the TIgGER inference process.

To help us filter out potential false positives, we inspected the mismatch frequencies of polymorphic positions. We took all sequences with the same allele annotation, including those with a suspected polymorphism, and calculated the frequencies of nucleotide mismatches at the polymorphic site in each individual. Finally, the individuals' mismatch frequencies for each novel polymorphism candidate were plotted (Figure 2). Mismatch frequency of 1.00 would correspond to an individual who was homozygous for a novel polymorphism candidate, i.e., all sequences contained the novel polymorphism and no other variant was detected. In theory, a heterozygous individual would be expected to have a mismatch frequency around 0.50, where around half of the sequences would contain the novel polymorphism and the rest of sequences would be identical to the germline reference (mismatch frequency = 0.00). To allow for potential copy number variation, we set a mismatch frequency cutoff value at 0.25, which should allow us to detect equal expression of up to four different variants of the same gene.

Only those novel allele candidates where at least one individual had a mismatch frequency above the cutoff value were considered true (Figure 2). Overall, 48 novel V allele candidates were above the mismatch cutoff; 25 in kappa and 23 in lambda genes.

Mismatch frequency pattern helps distinguish MiSeq errors from true polymorphisms

It is known that next-generation sequencing methods are not error proof and sequencing artefacts can be present in the sequencing data (Schirmer et al., 2016). When inferring germline alleles from receptor repertoires, sequencing artefacts can be mistaken for novel polymorphisms. For MiSeq in particular, A > C substitutions are known to be the most frequent substitution errors (Schirmer et al., 2015). Of all our inferred novel polymorphisms, 20 were A > C substitutions (Figure 1). We also identified four novel allele candidates

Table 1. Characteristics of suspected MiSeq errors

A > C polymorphism	Inference software	Mismatch frequency pattern	Repeat motif	Sanger-SEQ result
IGLV1-44*01_A207C	IgDiscover	Stack-like	No	Not validated
IGLV7-46*01_A213C	IgDiscover	Multimodal	Yes	Validated
IGLV10-54*01_A228C	TIgGER, IgDiscover	Multimodal	No	Validated

(IGLV2-11*01_T205C, IGLV7-46*01_A213C, IGLV3-19*01_A207G, IGLV1-44*01_T205C) that occurred in a repeat, i.e., the surrounding nucleotides were identical to the novel polymorphism (e.g., CCAC > CCCC). Such repeats were found to be more likely to occur due to MiSeq errors (Schirmer et al., 2015).

On top of that, we noticed a suspicious mismatch frequency pattern in some novel allele candidates (Figure 2). Most novel polymorphism candidates had a multimodal pattern where individuals were distributed in separate clusters according to their mismatch frequency value. However, in some cases, the mismatch frequencies of certain putative polymorphisms created a stack-like pattern, which in most cases did not pass the 0.25 cutoff value (Figure 2). Altogether, these observations made us suspicious whether the inferred polymorphisms could in fact result from MiSeq artefacts. To determine this, we chose three novel A > C polymorphism candidates for validation by Sanger sequencing: IGLV1-44*01_A207C, IGLV7-46*01_A213C, and IGLV10-54*01_A228C (Table 1). Of these, IGLV7-46*01_A213C polymorphism was located in a repeat motif and IGLV1-44*01_A207C has a stack-like mismatch frequency pattern. Two of these polymorphisms were found by IgDiscover only (IGLV1-44*01_A207C, IGLV7-46*01_A213C), and one was found by both TIgGER and IgDiscover (IGLV10-54*01_A228C).

We attempted to validate these three novel allele candidates by targeted amplification of the respective genes using genomic DNA (gDNA) of individuals in which the polymorphism was inferred. The PCR products were subsequently cloned into a bacterial vector and Sanger sequenced. Two of the selected candidates, namely, IGLV7-46*01_A213C and IGLV10-54*01_A228C, were successfully validated (Table 1). Amplification and Sanger sequencing of IGLV1-44 only revealed IGLV1-44*01 without any polymorphisms, despite choosing an individual with mismatch frequency above the 0.25 cutoff value. We attempted to validate this candidate from two additional individuals; however, we were not able to detect it in any of them. This led us to believe that IGLV1-44*01_A207C is most likely a sequencing artefact.

The difference between the validated and non-validated candidates was in their mismatch frequency pattern. IGLV7-46*01_A213C and IGLV10-54*01_A228C, which were both successfully validated, had a multimodal mismatch frequency pattern. In contrast, the novel polymorphism candidate IGLV1-44*01_A207C, which was not validated by Sanger sequencing, had a stack-like mismatch frequency pattern. Therefore, comparison of mismatch frequency patterns in a cohort seems to be useful for distinguishing MiSeq artefacts from true germline polymorphisms.

Validation of selected novel allele candidates reveals additional polymorphisms

We selected 11 novel allele candidates for validation (in addition to the three mentioned in the previous section). Validation was done by targeted amplification of the respective gene from gDNA of an individual with the suspected polymorphism. The PCR products were cloned into a pGEM-T Easy bacterial vector and Sanger sequenced using a universal T7 primer. Of the selected candidates, additional nine were successfully validated. Altogether, 11 of 14 selected novel allele candidates were successfully validated (Table S2). Fasta sequences of all validated polymorphisms can be found in Data File S1.

Attempts to validate IGLV3-25*03_G74T from two different individuals were inconclusive. At least one clone failed several Sanger sequencing runs, and/or the quality of the obtained sequence was low and contained Ns in the V-REGION. Although the reason for the failed or low-quality runs is not precisely known, there is a high possibility that this could be due to secondary structures, which might prevent amplification of the target alleles. Selecting a sequencing protocol for GC-rich sequences was slightly helpful but did not completely resolve the issue. Closer inspection of the novel allele candidate sequence revealed that the inferred polymorphism G74T in IGLV3-25*03 enables the formation of a hairpin loop, which would otherwise not be possible without this mutation.

We were unable to obtain the gDNA sequence of IGKV4-1*01_G72A. The individual, whose gDNA was used for validation, was heterozygous for this putative novel allele, and we were only able to amplify and sequence IGKV4-1*01. We suspect that polymorphisms in the primer-binding site might have prevented the amplification of the potential novel allele.

During our attempts to validate polymorphisms in the V-REGION by amplification and Sanger sequencing of gDNA, we also detected polymorphisms in other parts of the V gene. When inspecting the alignment of gDNA Sanger sequences to the reference sequences from the IMGT germline reference database (IMGT/GENE-DB), we noticed what appeared to be a 21-nt insertion in the promoter of IGKV1-8. This 21-nt stretch contains the decamer of the IGKV1 promoter. Although this fragment is absent in the reference sequence of IGKV1-8*01 in the IMGT/GENE-DB, all of our gDNA sequences corresponding to the IGKV1-8*01 allele contained this 21-nt segment. In addition to the V-REGION polymorphisms that were validated, the novel allele IGKV1-8*01_T31C was found to have more polymorphisms. One polymorphism was present in the promoter region (G > A), one in the 5'UTR (A > G), and another polymorphism (A > G) was found in the 3'end downstream of the V recombination signal sequence (V-RS) (Figure S1). The same polymorphisms that were found in the upstream and downstream of the V-REGION in IGKV1-8*01_T31C were also observed in the gDNA sequence of the novel allele IGKV1-8*01_G12A_T31C.

Relative abundance of light chain genes and alleles

To get a better overview of the light chain genes and alleles utilized in a naïve repertoire, we analyzed the relative usage of all kappa and lambda genes that were found in the cohort (Figure S2). As we previously described (Gidoni et al., 2019), the relative gene usage can serve as an indicator for double chromosome deletions. To infer such deletions a binomial test can be used. In genes that exhibit a multimodal distribution, the test checks whether individuals in the lower mod are significantly far from the distribution-derived threshold. For this, we utilized the binomial method that was used in Gidoni et al., (2019) and adjusted the minimum threshold for each gene. The minimum threshold was set as the usage closest to 0.001 for V and 0.005 for J. The difference in the cutoff value is due to the lower number of J genes compared with V genes. The individual gene minimal fraction for the binomial test was set as the point closest to the minimum cutoff. For this analysis we selected individuals who had a minimum of 2000 sequences and less than 3 mutations within the V region.

Observing the relative usage of the IGLJ genes indicated potential double chromosome deletions in two of them, IGLJ7 and IGLJ6 (Figure 3). The gene IGLJ7 is lowly expressed in the cohort, with a maximum relative usage of 0.28%. Even so, its relative usage in the cohort follows a bi-modal distribution. For eight individuals, the IGLJ7 usage was lower, which accounts for the lower mod of the distribution. Applying the binomial test for this gene resulted in a significant low adjusted p value (<0.01) for three individuals, which indicates a double chromosome deletion. A deletion for any J gene was never reported and no genomic validation was performed. Consequently, in this case the double chromosome deletion may be interpreted as deleted from the repertoire and not necessarily from the genomic locus itself. For the gene IGLJ6, all individual relative usage was below the minimum threshold (Figure 3), hence the test was undetermined. As the usage for most individuals was very close to zero, it might be speculated that this gene could be a pseudogene.

Approximately half of the lambda V genes are utilized more frequently than the others (mean above 0.025), and their relative usage varies greatly within the population (Figure 3). Three of the lambda V genes, IGLV5-52, IGLV3-22, and IGLV4-3, were detected in low frequency, the mean usage for these genes was lower than the initial threshold (0.001) and thus were declared unknown for the binomial test. One additional lambda V gene, IGLV3-12, was hardly detected (mean usage less than 0.0001) in the cohort and might be suspected to be a pseudogene.

As for the kappa locus, all IGKJ genes were detected in the cohort (Figure 4). In comparison with other Ig loci, the IGKV cluster is more complex. This region carries a duplication of a large part of the locus that carries many genes. These duplicated genes often share the same alleles, which causes calling multiple assignments while annotating the sequences. As a result, this affects our ability to correctly assess relative gene usage. This was corrected with the new annotations described above, which helped to obtain a better picture of the relative usage of this population. Several V genes exhibit a bimodal distribution, in some instances revealing double chromosome deletion such as IGKV2-29, yet in others, the IGKV2-29 gene usage

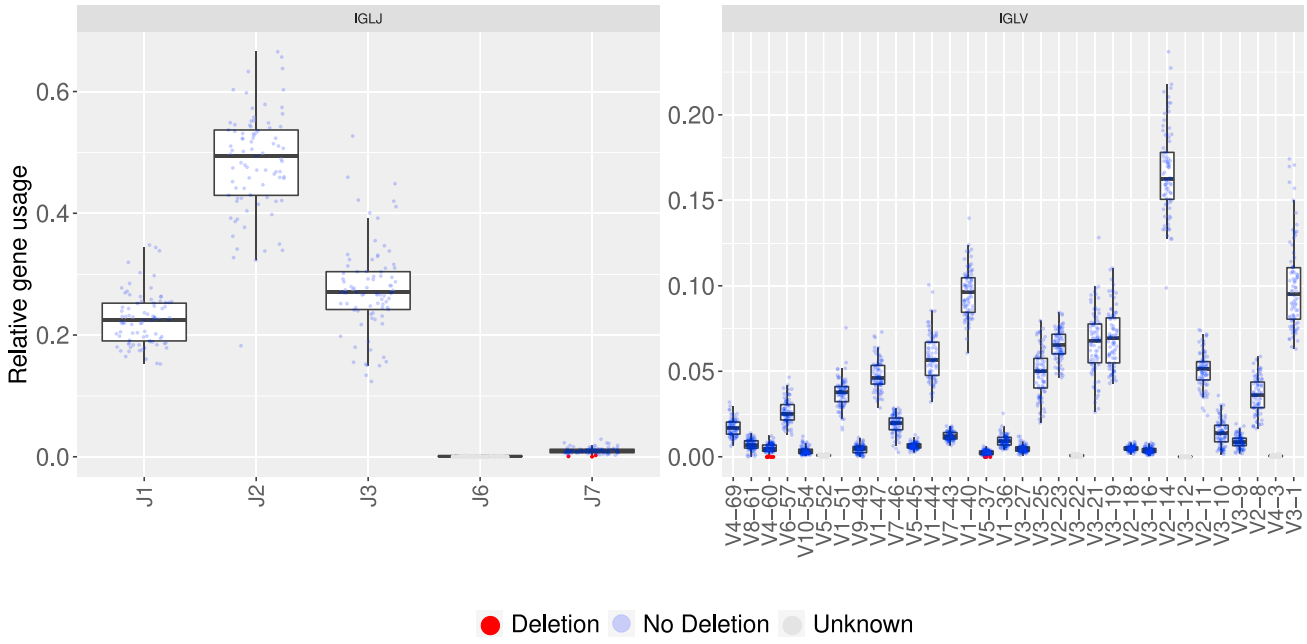


Figure 3. IGLJ6 and IGLV3-12 are suspected to be pseudogenes

These plots show the relative usage of lambda genes in individuals from our cohort. The x axis shows the different genes (IGLJ on the left panel, IGLV on the right), and the y axis shows the fraction of the relative gene usage. Each gene is represented by a box plot, whereas each individual is represented by a dot. The colors represent the presence of double chromosome deletion: red, deletion; light blue, no deletion; and dark gray, unknown. Only functional lambda V genes are shown in this plot, and all genes detected in the dataset and their respective usage are shown in Figure S2. For more detailed representation of IGLJ7 usage see also Figure S7.

is considered high (Figure 4). Same as in the lambda locus, several kappa V genes were detected in low frequency, below the minimal threshold and thus were marked as unknown in the test (IGKV2D-26, IGKV1D-17, IGKV3D-7).

Exploring the usage of IGKV1-5 (Figure S2) revealed an allele bias (Figure 5). After genotype inference, we detected that most individuals in our cohort carried the IGKV1-5*03 allele either in a homozygous form (no other alleles) or in combination with IGKV1-5*01 in heterozygotes (Figure 5A). The relative usage of this gene is distinct for individuals with different alleles. Individuals homozygous for the allele 01 have the highest relative usage of IGKV1-5, whereas this usage is lower for individuals carrying both alleles 01 and 03 (Figure 5B). Homozygous individuals carrying only IGKV1-5*03 have the lowest mean relative usage of IGKV1-5. The difference between the usage of IGKV1-5 alleles 01 and 03 in heterozygous individuals was consistently seen regardless of which IGKJ gene or allele it was recombined with.

Information revealed by haplotype inference

Haplotype inference can reveal patterns of chromosomal bias and single chromosome deletions. This was previously shown for the heavy chain in Gidoni et al. (2019), where several heavy chain genes exhibit a strong allele bias in their haplotypes. Similar to the heavy chain, the kappa locus expresses a single heterozygous J gene (IGKJ2). In contrast to that, no heterozygous J gene was observed on the lambda locus. A quarter of our cohort are heterozygous for IGKJ2. The ratio between the alleles varies but is sufficient for haplotype inference, where the minimum ratio is at least 30:70 between alleles. IGKJ2 has four alleles, and they tend to appear in heterozygous form with the alleles *01/*04 (18 individuals) and *01/*03 (6 individuals). Using the RAbHIT package (Peres et al., 2019) we inferred haplotypes with IGKJ2*01/*04 (see Figure S3). The haplotype map reinforced our suspicion of an allele bias in the IGKV1-5 gene. Individuals who are heterozygous for this gene exhibit a chromosomal bias. Three individuals (S11, S70, and S77) had a linkage between IGKV1-5*01 and IGKJ2*01, implying that both are located on the same chromosome. This adds to the evidence from Figure 5C of allele bias. The relative usage calculated for IGKV1-5 alleles depends on which J gene and allele they were recombined with. Heterozygous individuals for IGKV1-5 and IGKJ2 presented an allele bias, where the relative usage of allele IGKV1-5*01 was higher

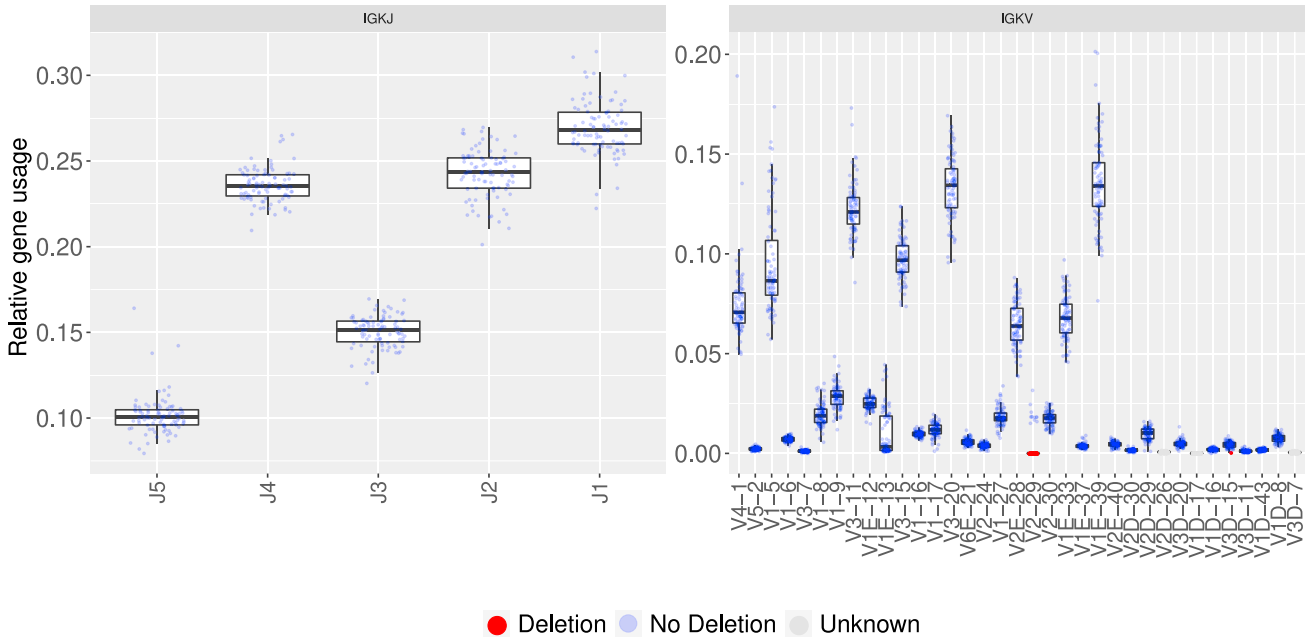


Figure 4. Kappa genes have varied relative usage

The panels in the figure show the relative usage of IGKJ (left) and IGKV (right) genes. The x axis shows the different genes, and the y axis shows the fraction of the relative gene usage. Each gene is represented by a box plot, whereas each individual is represented by a dot. The colors represent the presence of a suspected double chromosome deletion: red, deletion; light blue, no deletion; and dark gray, unknown. The "E" in the V gene names is used for duplicated genes that cannot be distinguished based on their V-REGION sequence. Only functional kappa V genes are shown in this plot, and all genes detected in the dataset and their respective usage are shown in Figure S2. More detailed usage distribution of IGKV genes with low usage is shown in Figure S8.

than IGKV1-5*03 allele. Homozygous individuals for IGKV1-5*03 have a relative usage around 0.5 (different color triangles in Figure 5C), which corresponds to an equal recombination frequency and what we observe in their haplotype (Figure S3).

Analysis of upstream sequences reveals alternative splicing

As previously shown (Mikocziova et al., 2020), upstream sequences of V genes, i.e., 5'UTR, L-PART1, and L-PART2, may also harbor polymorphisms. To characterize the upstream variants in our data, the upstream sequences from each individual and each gene/allele were clustered, and consensus sequences from clusters, which met the filtering criteria, were built. Another round of clustering was performed for consensus sequences obtained from all individuals, which resulted in a database of all upstream variants in our cohort. Details of the clustering steps and specific thresholds used are described in the methods section. Upstream variants are shown in Figures S4 and S5, and the nucleotide sequences are listed in a table in Data File S2.

In addition to regular upstream sequences, we also detected alternatively spliced transcripts in 11 kappa and 9 lambda V alleles (Figure S6). In these alternative transcripts, we observed partial intron retention (Figure 6B), which introduced a premature termination (stop) codon. These alternatively spliced transcripts were found in genes with low usage (Figures 3 and 4). We have higher confidence in variants that are present in multiple individuals or in a larger fraction of sequences (Figure 6A). However, since the total number of transcripts from lowly expressed genes is very small, our confidence in the inference of upstream variants from such genes is low.

DISCUSSION

Immunoglobulin light chain inference from repertoire sequencing is more challenging than the inference of heavy chain genes owing to various biological differences as well as challenges during the analysis process. Here, we have analyzed the light chain sequences from naïve BCR repertoires from a Norwegian cohort and addressed some of the computational challenges with the light chain allele inference. Our analysis revealed several previously unreported polymorphisms in the light chain V genes as well as alternatively spliced light chain transcripts.

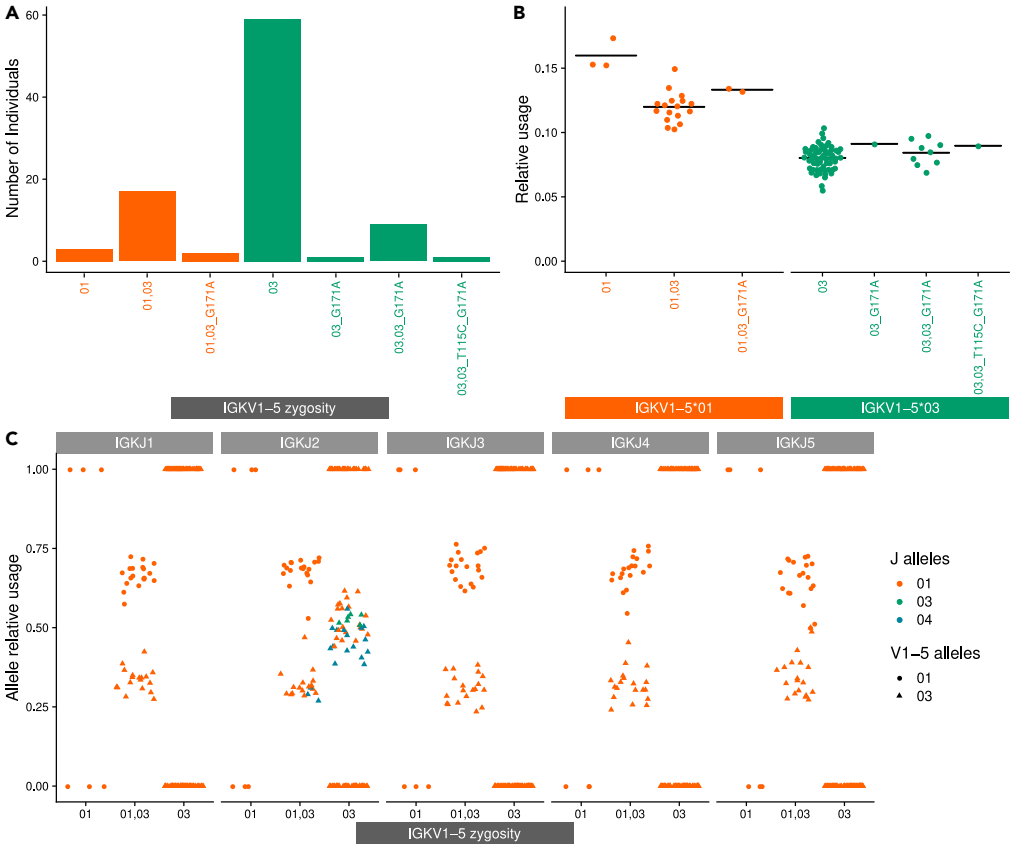


Figure 5. IGKV1-5 relative gene usage differs depending on the alleles in genotype

(A) The x axis is the combination of the IGKV1-5 alleles. The y axis is the individual count for each event. Orange represents allele combinations that include IGKV1-5*01, and green represents combinations without this allele.

(B) The y axis is the relative gene usage. The x axis is the combination of the IGKV1-5 alleles. Each dot represents an individual, orange represents allele combinations that include IGKV1-5*01, and green represents combinations without this allele.

(C) The x axis is the homozygous/heterozygous event without novel alleles. The y axis is the relative usage given a certain J gene. The color represents the J alleles, the shape represents the V alleles. Each two dots within a column represents an individuals' relative usage.

Inference of germline polymorphisms must be performed with great caution as high-throughput sequencing data are not 100% accurate due to PCR and sequencing errors. For example, A to C substitutions are one of the most common MiSeq errors, and substitution errors also occur frequently in certain repeats (Schirmer et al, 2015, 2016). In our analysis, we observed multiple A to C polymorphism candidates and we also identified a few novel allele candidates with a polymorphism within a repeat motif. Our validation attempts demonstrate that investigating the mismatch frequency of a polymorphic position for all individuals within a cohort can help with filtering out false positives.

Our results show preferential usage of certain genes and an allele bias in heterozygous individuals, as seen in the case of IGKV1-5. Similar bias was previously seen in the Ig heavy chain sequences from the same cohort (Gidoni et al., 2019). The underlying reasons for the bias remain unknown. We hypothesize that events happening during recombination and allelic exclusion could contribute to this bias. Further functional studies of the involved regulatory mechanism are needed to bring some clarity to this topic.

In addition to novel polymorphisms and biased allele usage, we also detected alternatively spliced mRNA of certain kappa genes by clustering their upstream sequences. These alternative transcripts retained a short fragment of the intron, which disrupted the reading frame and introduced a premature stop codon. Alternative splicing in kappa genes was already observed in 1980s and 1990s (Chou and Morrison, 1994;

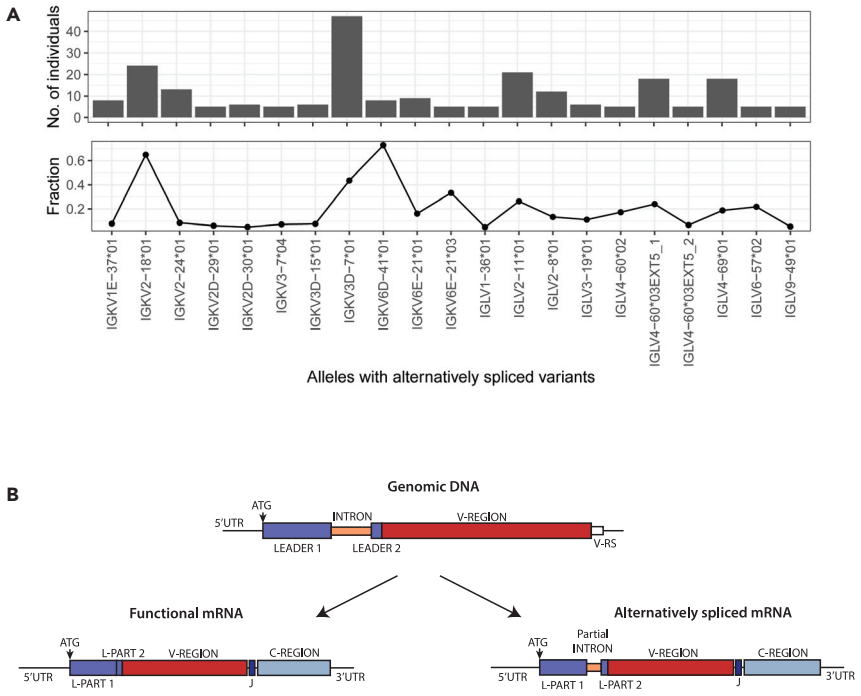


Figure 6. Alternatively spliced transcripts

(A) The x axis shows the different alleles for which alternative splice variants were detected. The y axis of the top plot represents the number of individuals in whom the variants were identified. The fraction represents the relative amount of the alternatively spliced sequences among the total sequences assigned to the respective germline allele.

(B) Schematic representation of alternative splicing of immunoglobulin transcripts.

Sikder et al., 1985), and Chou and Morrison (1994) even suggested that changes in the intron sequences might affect the expression of kappa genes.

It is known that transcripts with premature termination codons (PTCs) can be degraded by nonsense-mediated mRNA decay (NMD) pathway (Hug et al., 2016). The rate at which the degradation takes place seems to depend on the position of the PTC (Bühler et al., 2004). An *in vitro* study conducted by Bühler et al. (2004) showed that PTCs toward the 3′ end of the immunoglobulin mRNA get rapidly degraded by NMD, whereas PTCs at the 5′ end are not optimal substrates for NMD and do not get degraded as quickly. In our case, the alternative transcripts harbor PTCs closer toward the 5′ end, which could be the reason why we were able to detect them in the first place.

The question remains whether these supposedly unproductive transcripts can affect the expression of their respective genes. Unproductive splicing has been long believed to be a mechanism for regulating gene expression (Lareau et al., 2007; Lewis et al., 2003; Ni et al., 2007), and it can be observed across different animal kingdoms (Lareau and Brenner, 2015). However, the details of this mechanism remain largely unknown. There are not many studies that have looked at unproductive splicing of immunoglobulin genes. One older study showed a correlation between low levels of mRNA of kappa genes with stop codons and inefficient splicing (Lozano et al., 1994). In our data, the alternatively spliced mRNA was observed in genes with low relative usage. It might be possible that the unproductive transcript could somehow downregulate the transcription of its gene. However, the exact components and functional mechanisms involved in such regulation remain to be studied.

In summary, our study provides a strategy for the inference of germline polymorphisms optimized for the light chain Ig genes, which are frequently avoided in repertoire studies. We show that it is essential to filter out potential errors, for example, by exploring mismatch frequencies at polymorphic positions. For the first time, we show evidence of intron retention in alternatively spliced kappa and lambda transcripts in humans.

Limitations of the study

Although we report the relative abundance of different light chain genes and alleles in the data and use it as a proxy of their relative usage in the repertoire, our ability to correctly assess it is limited by current methods. It is conceivable that some alleles are not equally well amplified and therefore their true usage might not be well reflected in our dataset. In regard to the polymorphisms in the upstream sequences, owing to our filtering criteria, the number of variants might be underreported. Alternatively spliced sequences were present in very small numbers in the data, and to detect them, we had to relax some of our parameters, which might have introduced a certain extent of noise.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Data from human subjects
- ETHICS APPROVAL AND CONSENT TO PARTICIPATE
- METHOD DETAILS
  - Target gene amplification and cloning
  - Sanger sequencing and analysis
  - AIRR-seq data processing
  - Inference of upstream sequences
  - Haplotype and chromosomal deletion inference
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2021.103192.

ACKNOWLEDGMENTS

We would like to thank Knut E.A. Lundin for coordinating collection of blood samples of participating subjects and for being responsible for the ethical approval for the project. We would also like to thank Omri Snir and Ida Lindeman for providing gDNA samples for validation, Marie K. Johannesen and Bjørg Simonsen for technical assistance, and Aviv Omer for the helpful discussions. The authors are grateful to all study participants. This research was supported by Research Council of Norway through its Center of Excellence funding scheme [179573/V40]; South-Eastern Norway Regional Health Authority [2016113]; Stiftelsen KG Jebsen [SKGJ-MED-017 to L.M.S.]; ISF [832/16 to G.Y., M.G., and A.P.]; European Union's Horizon 2020 research and innovation program [825821]. The contents of this document are the sole responsibility of the iReceptor Plus Consortium and can under no circumstances be regarded as reflecting the position of the European Union.

AUTHOR CONTRIBUTIONS

L.M.S. and G.Y. conceived and designed the research; L.M.S., G.Y., and V.G. supervised the project; I.M. performed the experimental work; A.P., I.M., M.G., and G.Y. analyzed the data; I.M., A.P., L.M.S., G.Y., and V.G. wrote the paper. All authors edited the manuscript.

DECLARATION OF INTERESTS

V.G. declares advisory board positions in aiNET GmbH and Enpicom B.V. V.G. is also a consultant for Roche/Genentech. All remaining authors declare no conflict of interests.

Received: April 6, 2021
Revised: July 17, 2021
Accepted: September 27, 2021
Published: October 22, 2021

REFERENCES

Akbar, R., Robert, P.A., Pavlović, M., Jeliazkov, J.R., Snapkov, I., Slabodkin, A., Weber, C.R., Scheffer, L., Miho, E., Haff, I.H., et al. (2021). A compact vocabulary of paratope-epitope interactions enables predictability of antibody-antigen binding. Cell Rep *34*, 108856.

Alamyar, E., Duroux, P., Lefranc, M.-P., and Giudicelli, V. (2012). IMGT® tools for the nucleotide analysis of immunoglobulin (IG) and T cell receptor (TR) V-(D)-J repertoires, polymorphisms, and IG mutations: IMGT/V-QUEST and IMGT/HighV-QUEST for NGS. In Immunogenetics: Methods and Applications in Clinical Practice, Methods in Molecular Biology, F.T. Christiansen and B.D. Tait, eds. (Humana Press), pp. 569–604.

Avnir, Y., Tallarico, A.S., Zhu, Q., Bennett, A.S., Connelly, G., Sheehan, J., Sui, J., Fahmy, A., Huang, C.-Y., Cadwell, G., et al. (2014). Molecular signatures of hemagglutinin stem-directed heterosubtypic human neutralizing antibodies against influenza A viruses. Plos Pathog. *10*, e1004103.

Avnir, Y., Watson, C.T., Glanville, J., Peterson, E.C., Tallarico, A.S., Bennett, A.S., Qin, K., Fu, Y., Huang, C.-Y., Beigel, J.H., et al. (2016). IGHV1-69 polymorphism modulates anti-influenza antibody repertoires, correlates with IGHV utilization shifts and varies by ethnicity. Sci. Rep. *6*, 20842.

Bühler, M., Paillusson, A., and Mühlemann, O. (2004). Efficient downregulation of immunoglobulin μ mRNA with premature translation-termination codons requires the 5′-half of the VDJ exon. Nucl. Acids Res. *32*, 3304–3315.

Chi, X., Li, Y., and Qiu, X. (2020). V(D)J recombination, somatic hypermutation and class switch recombination of immunoglobulins: mechanism and regulation. Immunology *160*, 233–247.

Chou, C.L., and Morrison, S.L. (1994). Intron sequences determine the expression of kappa light chain genes. Mol. Immunol. *31*, 99–107.

Collins, A.M., and Watson, C.T. (2018). Immunoglobulin light chain gene rearrangements, receptor editing and the development of a self-tolerant antibody repertoire. Front. Immunol. *9*. https://doi.org/10.3389/fimmu.2018.02249.

Collins, A.M., Yaari, G., Shepherd, A.J., Lees, W., and Watson, C.T. (2020). Germline immunoglobulin genes: disease susceptibility genes hidden in plain sight? Curr. Opin. Syst. Biol. https://doi.org/10.1016/j.coisb.2020.10.011.

Corcoran, M.M., Phad, G.E., Bernat, N.V., Stahl-Hennig, C., Sumida, N., Persson, M.A.A., Martin, M., and Hedestam, G.B.K. (2016). Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity. Nat. Commun. *7*, 13642.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucl. Acids Res. *32*, 1792–1797.

Gadala-Maria, D., Gidoni, M., Marquez, S., Vander Heiden, J.A., Kos, J.T., Watson, C.T., O'Connor, K.C., Yaari, G., and Kleinstein, S.H. (2019). Identification of subject-specific immunoglobulin alleles from expressed repertoire sequencing data. Front. Immunol. *10*. https://doi.org/10.3389/fimmu.2019.00129.

Gadala-Maria, D., Yaari, G., Uduman, M., and Kleinstein, S.H. (2015). Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles. Proc. Natl. Acad. Sci. U S A *112*, E862–E870.

Gidoni, M., Snir, O., Peres, A., Polak, P., Lindeman, I., Mikocziova, I., Sarna, V.K., Lundin, K.E.A., Clouser, C., Vigneault, F., et al. (2019). Mosaic deletion patterns of the human antibody heavy chain gene locus shown by Bayesian haplotyping. Nat. Commun. *10*, 628.

Giudicelli, V., Chaume, D., and Lefranc, M.-P. (2005). IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. Nucl. Acids Res. *33*, D256–D261.

Glanville, J., Kuo, T.C., Büdingen, H.-C., Guey, L., Berka, J., Sundar, P.D., Huerta, G., Mehta, G.R., Oksenberg, J.R., Hauser, S.L., et al. (2011). Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. Proc. Natl. Acad. Sci. U S A *108*, 20066–20071.

Hug, N., Longman, D., and Cáceres, J.F. (2016). Mechanism and regulation of the nonsense-mediated decay pathway. Nucl. Acids Res. *44*, 1483–1495.

Johnson, T.A., Mashimo, Y., Wu, J.-Y., Yoon, D., Hata, A., Kubo, M., Takahashi, A., Tsunoda, T., Ozaki, K., Tanaka, T., et al. (2020). Association of an IGHV3-66 gene variant with Kawasaki disease. J. Hum. Genet. 1–15.

Lareau, L.F., and Brenner, S.E. (2015). Regulation of splicing factors by alternative splicing and NMD is conserved between kingdoms yet evolutionarily flexible. Mol. Biol. Evol. *32*, 1072–1079.

Lareau, L.F., Inada, M., Green, R.E., Wengrod, J.C., and Brenner, S.E. (2007). Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. Nature *446*, 926–929.

Lewis, B.P., Green, R.E., and Brenner, S.E. (2003). Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc. Natl. Acad. Sci. U S A *100*, 189–192.

Lozano, F., Maertzdorf, B., Pannell, R., and Milstein, C. (1994). Low cytoplasmic mRNA levels of immunoglobulin kappa light chain genes containing nonsense codons correlate with inefficient splicing. EMBO J. *13*, 4617–4622.

Luo, S., Yu, J.A., Li, H., and Song, Y.S. (2019). Worldwide genetic variation of the IGHV and TRBV immune receptor gene families in humans. Life Sci. Alliance *2*. https://doi.org/10.26508/lsa.201800221.

Madeira, F., Park, Y.M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A.R.N., Potter, S.C., Finn, R.D., and Lopez, R. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucl. Acids Res. *47*, W636–W641.

McBride, O.W., Heiter, P.A., Hollis, G.F., Swan, D., Otey, M.C., and Leder, P. (1982). Chromosomal location of human kappa and lambda immunoglobulin light chain constant region genes. J. Exp. Med. *155*, 1480–1490.

Mikocziova, I., Gidoni, M., Lindeman, I., Peres, A., Snir, O., Yaari, G., and Sollid, L.M. (2020). Polymorphisms in human immunoglobulin heavy chain variable genes and their upstream regions. Nucl. Acids Res. *48*, 5499–5510.

Mikocziova, I., Greiff, V., and Sollid, L.M. (2021). Immunoglobulin germline gene variation and its impact on human disease. Genes Immun. 1–13.

Ni, J.Z., Grate, L., Donohue, J.P., Preston, C., Nobida, N., O'Brien, G., Shiue, L., Clark, T.A., Blume, J.E., and Ares, M. (2007). Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. Genes Dev. *21*, 708–718.

Peres, A., Gidoni, M., Polak, P., and Yaari, G. (2019). RAbHIT: R Antibody Haplotype Inference Tool. Bioinformatics *35* (22), 4840–4842.

Ralph, D.K., and Matsen (IV), F.A. (2019). Per-sample immunoglobulin germline inference from B cell receptor deep sequencing data. Plos Comput. Biol. *15*, e1007133.

Rubelt, F., Bolen, C.R., McGuire, H.M., Heiden, J.A.V., Gadala-Maria, D., Levin, M., Euskirchen, M., Mamedov, M.R., Swan, G.E., et al. (2016). Individual heritable differences result in unique cell lymphocyte receptor repertoires of naïve and antigen-experienced cells. Nat. Commun. *7*, 11112.

Schatz, D.G. (2004). V(D)J recombination. Immunol. Rev. *200*, 5–11.

Scheepers, C., Shrestha, R.K., Lambson, B.E., Jackson, K.J.L., Wright, I.A., Naicker, D., Goosen, M., Berrie, L., Ismail, A., Garrett, N., et al. (2015). Ability to develop broadly neutralizing HIV-1 antibodies is not restricted by the germline Ig gene repertoire. J. Immunol. *194*, 4371–4378.

Schirmer, M., D'Amore, R., Ijaz, U.Z., Hall, N., and Quince, C. (2016). Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. BMC Bioinform. *17*, 125.

Schirmer, M., Ijaz, U.Z., D'Amore, R., Hall, N., Sloan, W.T., and Quince, C. (2015). Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucl. Acids Res. *43*, e37.

Sela-Culang, I., Kunik, V., and Ofran, Y. (2013). The structural basis of antibody-antigen recognition. Front. Immunol. *4*. https://doi.org/10.3389/fimmu.2013.00302.

Sikder, S.K., Kabat, E.A., and Morrison, S.L. (1985). Alternative splicing patterns in an aberrantly rearranged immunoglobulin kappa-light-chain gene. Proc. Natl. Acad. Sci. U S A *82*, 4045–4049.

Slabodkin, A., Chernigovskaya, M., Mikocziova, I., Akbar, R., Scheffer, L., Pavlović, M., Bashour, H.,

Snapkov, I., Mehta, B.B., Weber, C.R., et al. (2021). Individualized VDJ recombination predisposes the available Ig sequence space. bioRxiv. https://doi.org/10.1101/2021.04.19.440409.

Tan, J., Sack, B.K., Oyen, D., Zenklusen, I., Piccoli, L., Barbieri, S., Foglierini, M., Fregni, C.S., Marcandalli, J., Jongo, S., et al. (2018). A public antibody lineage that potently inhibits malaria infection through dual binding to the circumsporozoite protein. Nat. Med. *24*, 401–407.

Thörnqvist, L., and Ohlin, M. (2018). The functional 3′-end of immunoglobulin heavy chain variable (IGHV) genes. Mol. Immunol. *96*, 61–68.

Tobin, G. (2005). The immunoglobulin genes and chronic lymphocytic leukemia (CLL). Ups. J. Med. Sci. *110*, 97–114.

Vander Heiden, J.A., Yaari, G., Uduman, M., Stern, J.N.H., O'Connor, K.C., Hafler, D.A., Vigneault, F., and Kleinstein, S.H. (2014). pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. Bioinformatics *30*, 1930–1932.

Vázquez Bernat, N., Corcoran, M., Hardt, U., Kaduk, M., Phad, G.E., Martin, M., and Karlsson Hedestam, G.B. (2019). High-quality library preparation for NGS-based immunoglobulin germline gene inference and repertoire expression analysis. Front. Immunol. *10*, 660.

Vencovský, J., Zď árský, E., Moyes, S.P., Hajeer, A., Ruzicková, S., Cimburek, Z., Ollier, W.E., Maini, R.N., and Mageed, R.A. (2002). Polymorphism in the immunoglobulin VH gene V1-69 affects susceptibility to rheumatoid arthritis in subjects lacking the HLA-DRB1 shared epitope. Rheumatology *41*, 401–410.

Watson, C.T., Glanville, J., and Marasco, W.A. (2017). The individual and population genetics of antibody immunity. Trends Immunol. *38*, 459–470.

Watson, C.T., Steinberg, K.M., Graves, T.A., Warren, R.L., Malig, M., Schein, J., Wilson, R.K., Holt, R.A., Eichler, E.E., and Breden, F. (2015). Sequencing of the human IG light chain loci from a hydatidiform mole BAC library reveals locus-specific signatures of genetic diversity. Genes Immun. *16*, 24–34.

Xochelli, A., Agathangelidis, A., Kavakiotis, I., Minga, E., Sutton, L.A., Baliakas, P., Chouvarda, I., Giudicelli, V., Vlahavas, I., Maglaveras, N., et al. (2015). Immunoglobulin heavy variable (IGHV) genes and alleles: new entities, new names and implications for research and prognostication in chronic lymphocytic leukaemia. Immunogenetics *67*, 61–66.

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., and Madden, T.L. (2012). Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. BMC Bioinform. *13*, 134.

Ye, J., Ma, N., Madden, T.L., and Ostell, J.M. (2013). IgBLAST: an immunoglobulin variable domain sequence analysis tool. Nucl. Acids Res. *41*, W34–W40.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and virus strains		
E.coli XL10 $CaCl_2$-competent cells	In-house made	N/A
Biological samples		
Genomic DNA samples from 100 individuals recruited in Norway	Gidoni et al., 2019; Mikocziova et al., 2020	N/A
Chemicals, peptides, and recombinant proteins		
Q5® Hot Start High-Fidelity DNA Polymerase	New England Biolabs	Cat#:M0493S
DNA Polymerase I, Large (Klenow) Fragment	New England Biolabs	Cat#:M0210S
NEBNext® dA-Tailing Reaction Buffer	New England Biolabs	Cat#:B6059S
pGEM®-T Easy vector	Promega	Cat#:A1360
X-Gal (5-bromo-4-chloro-3-indolyl-β-d-galactopyranoside)	Promega	Cat#: V3941
IPTG	Sigma-Aldrich	Cat#: I5502
Critical commercial assays		
Monarch® PCR & DNA Cleanup Kit (5 μg)	New England Biolabs	Cat#:T1030S
Monarch® DNA Gel Extraction Kit	New England Biolabs	Cat#:T1020S
Monarch® Plasmid Miniprep Kit	New England Biolabs	Cat#:T1010L
Deposited data		
Raw naïve Ig repertoire data	Gidoni et al., 2019	ENA: PRJEB26509 (ERP108501)
Validated gDNA sequences	This paper	GenBank: MW316667 – MW316678
Validated IGKV1-8_G12A_T31C gDNA sequence	This paper	GenBank: MW316669, New IUIS allele name: IGKV1-8*02
Validated IGKV1-8_T31C gDNA sequence	This paper	GenBank: MW316671, New IUIS allele name: IGKV1-8*03
Validated IGKV1-8_01 gDNA sequence (difference in the promoter)	This paper	GenBank: MW316672
Validated IGKV1-5_G171A gDNA sequence	This paper	GenBank: MW316670, New IUIS allele name: IGKV1-5*04
Validated IGKV1-5_T115C_G171A gDNA sequence		GenBank: MW316676, New IUIS allele name: IGKV1-5*05
Validated IGKV4-1_A318G gDNA sequence		GenBank: MW316673, New IUIS allele name: IGKV4-1*02
Validated IGLV1-44_C146G gDNA sequence		GenBank: MW316667, New IUIS allele name: IGLV1-44*02

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Validated IGLV1-44_C215T gDNA sequence		GenBank: MW316668, New IUIS allele name: IGLV1-44*03
Validated IGLV2-14_T213C_T214C_A220G gDNA sequence		GenBank: MW316674, New IUIS allele name: IGLV2-14*05
Validated IGLV7-46_A213C gDNA sequence		GenBank: MW316675, New IUIS allele name: IGLV7-46*04
Validated IGLV6-57_C21G gDNA sequence		GenBank: MW316677
Validated IGLV10-54_A228C gDNA sequence		GenBank: MW316678
Oligonucleotides		
IGLV1-44_Fwd TCAGGGCTCACAACTGTGTT	This paper	N/A
IGLV1-44_Rev GACCCTGTGTCCTAAGCTGC	This paper	N/A
IGKV1-5_Fwd TGCATGTTCCCAGAGCACAA	This paper	N/A
IGKV1-5_Rev AGTCCAGCTGAAGCCATAAAC	This paper	N/A
IGKV1-8_Fwd TCCAAATAATCCCCATGTGCCA	This paper	N/A
IGKV1-8_Rev TCCCCCTCTACCAACACCAT	This paper	N/A
IGKV4-1_Fwd TTCTACGATGCACAAGGCGT	This paper	N/A
IGKV4-1_Rev CCCAACACACAGAGGAAGCA	This paper	N/A
Primers for IGLV2-14, IGLV3-25, IGLV6-57, IGLV7-46, and IGLV10-54 are listed in the Table S1	Vázquez Bernat et al., 2019	N/A
Software and algorithms		
R v3.6	R Core Team	https://www.R-project.org/
pRESTO	Vander Heiden et al., 2014	https://presto.readthedocs.io/en/stable/
IgBLAST	Ye et al., 2013	https://ncbi.github.io/igblast/
TIgGER v 1.0.0	Gadala-Maria et al., 2015; 2019	https://tigger.readthedocs.io/en/stable/
IgDiscover v0.12	Corcoran et al., 2016	https://github.com/NBISweden/IgDiscover
RAbHIT v0.1.5	Peres et al., 2019	https://bitbucket.org/yaarilab/rabhit
Processing pipeline	This paper	https://bitbucket.org/yaarilab/processpipeline
5'UTR analysis	Modified from Mikocziova et al., 2020	https://bitbucket.org/yaarilab/cluster_5utr

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Ludvig M. Sollid, University of Oslo (l.m.sollid@medisin.uio.no).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The raw naïve Ig repertoire data comes from a previously published study (Gidoni et al., 2019) and is available at the European Nucleotide Archive (ENA: PRJEB26509, Secondary Study Accession: ERP108501).

The DNA sequences of the validated novel alleles obtained by Sanger sequencing were submitted to GenBank (GenBank: MW316667–MW316678) and are and are publicly available as of the date of publication. Their accession numbers as well as their new allele names assigned by the IUIS Immunoglobulins (IG), T cell Receptors (TR) and Major Histocompatibility (MH) Nomenclature Sub-Committee are specified in the Key Resources Table. The DNA sequences are also provided as a supplementary file Data File S1.

A summary of all upstream sequences is available as a table in the supplementary file Data File S2. The associated README.txt document provides a description of the columns in the table.

The code used for the processing pipeline (https://bitbucket.org/yaarilab/processpipeline); inference of haplotypes, deletions, and relative usage (https://bitbucket.org/yaarilab/rabhit); and analysis of upstream sequences (https://bitbucket.org/yaarilab/cluster_5utr) is publicly available as of the date of publication.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Data from human subjects

The raw naïve Ig repertoire data comes from a previously published study (Gidoni et al., 2019) and is available at the European Nucleotide Archive (ENA) under the accession number PRJEB26509 (ERP108501). The data contains naïve BCR repertoires from a Norwegian cohort of 100 individuals that were sequenced on the Illumina MiSeq platform (2x 300 bp PE).

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Data and sample material used in this study is covered by the approval of the Regional Ethical Committee (projects REK 2010/2720 and REK 2011/2472, project leader Knut E. A. Lundin). All participants gave written informed consent.

METHOD DETAILS

Target gene amplification and cloning

The target genes were amplified from non-B cell gDNA using gene-specific primers. The primer sequences used for amplification of IGLV2-14, IGLV3-25, IGLV6-57, IGLV7-46, and IGLV10-54 were taken from (Vázquez Bernat et al., 2019). Primers for IGKV1-5, IGKV1-8, IGKV4-1 and IGLV1-44 were designed using Primer-BLAST (Ye et al., 2012) with the following settings: Min length = 700, Max length = 1100; and Homo sapiens RefSeq representative genome as a template. The remaining parameters were left as default. All primers were synthesized by Eurogentec (RP-cartridge purification). The nucleotide sequences of the primers are listed in Table S1.

The cloning was performed as previously described (Mikocziova et al., 2020). The target genes were amplified using the above described primers with the Q5® Hot Start High-Fidelity DNA Polymerase (NEB). Touch-down PCR protocol was used to avoid possible off-target amplification. Following amplification, the PCR products were analyzed by gel electrophoresis (1% agarose, 100 V, 45 min) and the bands were excised from the gel. DNA from the excised fragments was extracted using the Monarch® DNA Gel

Extraction Kit (NEB). Before cloning, the PCR products were then A-tailed using the Klenow fragment (DNA Polymerase I Large (Klenow) Fragment, NEB) and the NEBNext® dA-Tailing Reaction Buffer (NEB). The A-tailed PCR products were cleaned with Monarch® PCR & DNA Cleanup Kit (5 μg) (NEB) and ligated into pGEM®-T Easy vector (Promega) using 1:3 molar vector:insert ratio. The manufacturer's protocol was followed. Subsequently, 4 μl of the ligation reaction were used to transform 90 μl XL10 $CaCl_2$-competent cells. The transformed cells were plated onto LB_{amp} 50 μg/ml plates coated with IPTG/X-Gal (40 μl 100 mM IPTG + 16 μl 50 mg/ml X-Gal). From each plate, 4 white colonies were picked following an overnight incubation at 37°C. The picked colonies were cultured in suspension at 37°C for up to 14 h, and the plasmid DNA was extracted using the Monarch® Plasmid Miniprep Kit (NEB). The presence of the correct inserts was done by performing a PCR reaction using the same primers as for the initial target amplification and using the corresponding plasmid DNA as a template.

Sanger sequencing and analysis

Purified plasmid constructs were sent for Supremerun Sanger sequencing with a universal T7 primer to Eurofins/GATC. For problematic samples, the Supremerun option for GC-rich sequences was selected. The obtained sequences were preprocessed by trimming low quality ends and masking primers. Such pre-processed sequences were subsequently annotated using IMGT/HighV-QUEST (Alamyar et al., 2012). To compare the upstream sequences and introns, the obtained Sanger sequences were aligned to their respective reference sequence using MUSCLE (Edgar, 2004; Madeira et al., 2019). The reference germline immunoglobulin gene sequences were obtained from IMGT/GENE-DB (Giudicelli et al., 2005).

AIRR-seq data processing

pRESTO (Vander Heiden et al., 2014) was used for pre-processing the data as described before (Gidoni et al., 2019; Mikocziova et al., 2020). IgBLAST 1.14.0 (Ye et al., 2013) was applied to annotate the pre-processed data with a modified IMGT germline reference from August 2020 and modified parameters previously described (Mikocziova et al., 2020). Within the kappa loci, duplicated V genes that shared an identical allele were grouped together under a new name annotated with an E, and the alleles of the duplicated genes were added under a consecutive number. Several of the germline reference V genes are short at their 3′ and 5′ end, to better align these sequences we artificially extended the reference with collapsed ends version from the other known alleles of the same gene.

To strengthen the novel alleles inference, two tools were used: TIgGER v1.0.0 (Gadala-Maria et al, 2015, 2019) and IgDiscover v0.12 (Corcoran et al., 2016). The tools parameters were modified to lower the thresholds for novel allele discovery. For TIgGER all parameters except 'y_intercept' and 'alpha' were set to the minimum allowed, and for IgDiscover only the 'discover' step was applied. We filtered the inferences to allow up to two potential novel alleles for each gene, those with the highest exact match. The IGLV germline reference is longer than the IGHV, hence we had to adapt the defaulted TIgGER position range inference. Each gene had a custom 3′ max position range. For consistency in the final potential novel allele filtration, the genotype was inferred for both tools solely using TIgGER. For further analysis novel alleles were filtered by taking only genes with IK confident scores larger than 10.

Inference of upstream sequences

Analysis of upstream sequences (5′UTR, L-PART1 and L-PART2) was performed as previously described (Mikocziova et al., 2020). Briefly, the sequences were trimmed at the 5′ end of the V-REGION and the resulting sequences were collapsed based on their v_call annotations. The sequences were trimmed to a shared position and short 5′UTR sequences were removed based on frequency length above 0.05. Clusters and consensus sequences for each allele were constructed using ClusterSets.py (–ident 0.999, –length 0.5) and BuildConsensus.py (–freq 0.6) from pRESTO (Vander Heiden et al., 2014). The clusters with frequency below 0.01 and below 5 sequences were filtered out. For each allele a consensus sequence was constructed with ClusterSets.py (–ident 1.0, –length 1.0) and BuildConsensus.py (–freq 0.6). Lastly, sequences shared between the different individuals in the cohort were collapsed.

Haplotype and chromosomal deletion inference

Inference of haplotype and chromosomal deletion events for the light chain loci is an expansion of the methods previously described (Gidoni et al., 2019). Double chromosome deletion detection is based on variation of the relative gene usage within a population. When the relative gene usage of certain individuals

is much lower than the rest of the population a binomial test can assess whether a deletion event is present. The binomial test parameters are the sequences mapped to a certain gene (*X*), the total number of sequences (*N*), and the lowest relative frequency of this gene among candidates with relative frequencies larger than 0.001 and without a deletion event (*P*). To assert the p value for each V or J gene from the light chain loci, we calculated the relative gene usage of individuals that had more than 2000 sequences. For V gene deletion the candidate frequency threshold was set to 0.001 and for J to 0.005. Genes which had a mean usage lower than the initial threshold were not considered for the binomial test. Haplotype inference was performed using RAbHIT (Peres et al., 2019) package which utilizes a Bayesian framework based on a binomial likelihood. The package's haplotype inference function was modified to infer based heterozygous genes from the light chain loci. Heterozygosity was determined in a sufficient ratio of at least 30:70 between the alleles of IGKJ2, which allowed to infer haplotypes for heterozygous individuals. The same as the heavy chain, single chromosome deletion events were called when the "unknown" call of an allele in a certain chromosome that had a Bayes factor was larger than 1000.

QUANTIFICATION AND STATISTICAL ANALYSIS

Only samples with at least 2000 unique sequences were included in the study. To assess the double chromosome deletion a binomial test was applied as explained in the methods (section Haplotype and chromosomal deletion inference). Statistical analysis, genotype and haplotype inferences were performed using R v3.6, TIgGER v1.0.0 (Gadala-Maria et al., 2015, 2019), IgDiscover v0.12 (Corcoran et al., 2016), and RAbHIT v0.1.5 (Peres et al., 2019). All p values were adjusted for multiple hypotheses with Benjamini-Hochberg correction and the significance level was set to 0.01.